PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-132453
(43)Date of publication of application : 10.05.2002

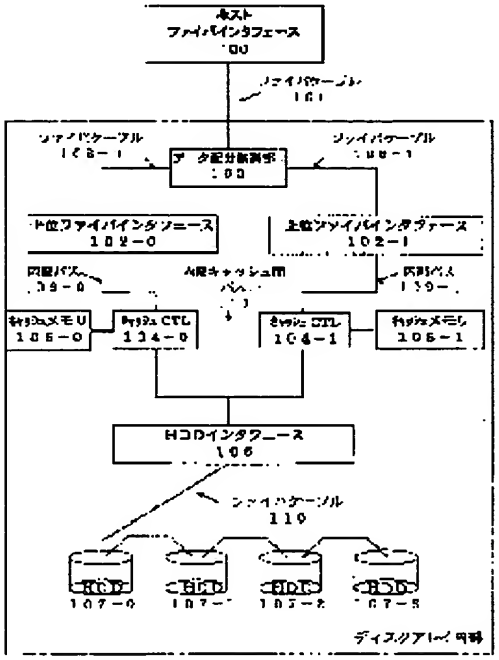
(51)Int.Cl. G06F 3/06
G06F 12/08

(21)Application number : 2000-321280
(22)Date of filing : 20.10.2000

(71)Applicant : HITACHI LTD
(72)Inventor : KYO SHIYOUMEI
YAMANASHI AKIRA
YAGISAWA IKUYA

(54) DISK ARRAY SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To realize a disk array having a high internal transfer capacity corresponding to a host computer with a high-speed fiber interface.
SOLUTION: This disk array system is equipped with plural upper fiber interfaces 102 that receive/send data from/to a host fiber interface 100, a cache memory 105 connected to the plural upper fiber interfaces, plural disk storage means 107 which data is written on or read from, and a disk drive interface 106 for controlling the storage means, and each of the plural upper fiber interfaces has a cache memory. Moreover, in the disk array system, a data distribution controlling part 103 is provided between the host fiber interface and the upper fiber interfaces.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－132453
（P2002－132453A）
(43)公開日　平成14年５月10日(2002.5.10)

(51)Int.Cl.[7]		識別記号	FI		テーマコード(参考)
Ｇ０６Ｆ	3/06	３０１	Ｇ０６Ｆ	3/06	３０１Ｕ　５Ｂ００５
					３０１Ｔ　５Ｂ０６５
		５４０			５４０
	12/08	５１１		12/08	５１１Ｚ
		５５７			５５７

審査請求　未請求　請求項の数８　ＯＬ　（全 12 頁）

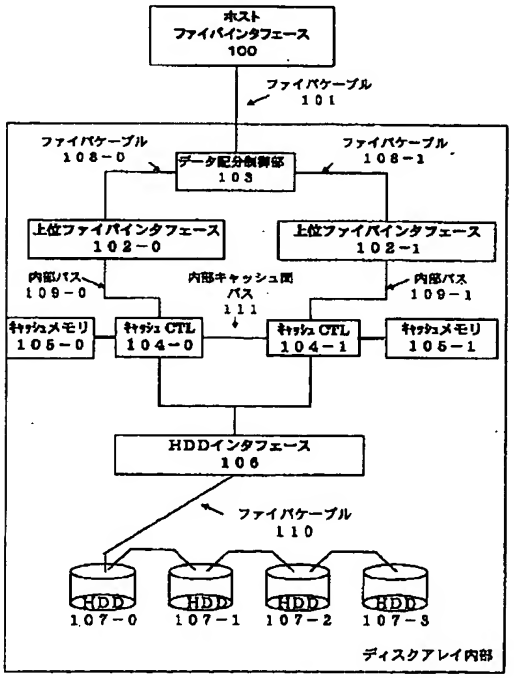
(21)出願番号　特願2000－321280(P2000－321280)
(22)出願日　平成12年10月20日(2000.10.20)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　姜　小明
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内
(72)発明者　山梨　晃
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内
(74)代理人　100093492
弁理士　鈴木　市郎　（外１名）

最終頁に続く

(54)【発明の名称】　ディスクアレイシステム

(57)【要約】
【課題】　高速ファイバインタフェースを持つホストコンピュータに対応した高い内部転送能力を有するディスクアレイシステムを実現すること。
【解決手段】ホストファイバインタフェース１００からのデータを授受する複数の上位ファイバインタフェース１０２と、前記複数上位ファイバインタフェースに接続されたキャッシュメモリ１０５と、データを書き込み又は読み出す複数のディスク記憶手段１０７と、前記記憶手段を制御するディスクドライブインタフェース１０６と、を備えたディスクアレイシステムであって、前記複数の上位ファイバインタフェースの各上位ファイバインタフェースに対応してそれぞれキャッシュメモリを備えている。また、前記ディスクアレイシステムにおいて、ホストファイバインタフェースと上位ファイバインタフェースとの間にデータ配分制御部１０３を設ける。



1

【特許請求の範囲】

【請求項１】　ホストファイバインタフェースからのデータを授受する複数の上位ファイバインタフェースと、前記複数上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライブインタフェースと、を備えたディスクアレイシステムであって、

前記複数の上位ファイバインタフェースの各上位ファイバインタフェースに対応してそれぞれキャッシュメモリを備えていることを特徴とするディスクアレイシステム。

【請求項２】　請求項１に記載のディスクアレイシステムにおいて、

前記ホストファイバインタフェースと前記上位ファイバインタフェースとの間に設けられたデータ配分制御部は複数のデータバッファを備え、

前記データ配分制御部は、前記キャッシュメモリの空き容量の管理を行い、前記ホストファイバインタフェースからのデータを前記データバッファへ格納し、又は前記データバッファのバッファデータを前記キャッシュメモリに転送する、機能を有することを特徴とするディスクアレイシステム。

【請求項３】　請求項２に記載のディスクアレイシステムにおいて、

前記データ配分制御部はキャッシュメモリ内のキャッシュ管理テーブルのコピーを有することを特徴とするディスクアレイシステム。

【請求項４】　請求項１に記載のディスクアレイシステムにおいて、

前記各上位ファイバインタフェースに対応してそれぞれ備えられたキャッシュメモリ同士間でデータ及び管理情報を授受するための通信バスを有することを特徴とするディスクアレイシステム。

【請求項５】　ホストファイバインタフェースからのデータを授受する複数の上位ファイバインタフェースと、前記複数上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライブインタフェースと、を備えたディスクアレイシステムであって、

前記ホストファイバインタフェースと前記上位ファイバインタフェースとの間にデータ配分制御部を設け、

前記データ配分制御部は、前記ホストファイバインタフェースからのデータを複数に分割して前記複数の上位ファイバインタフェースに供給するとともに、前記複数上位ファイバインタフェースを介して前記分割されたデータを並列に複数キャッシュメモリに同時に転送することを特徴とするディスクアレイシステム。

【請求項６】　請求項５に記載のディスクアレイシステムにおいて、

2

前記データ配分制御部は複数のデータバッファを備え、

前記データ配分制御部は、前記キャッシュメモリの空き容量の管理を行い、前記ホストファイバインタフェースからのデータを前記データバッファへ格納し、又は前記データバッファのバッファデータを前記キャッシュメモリに転送する、機能を有することを特徴とするディスクアレイシステム。

【請求項７】　請求項５に記載のディスクアレイシステムにおいて、

前記データ配分制御部はキャッシュメモリ内のキャッシュ管理テーブルのコピーを有することを特徴とするディスクアレイシステム。

【請求項８】　請求項５に記載のディスクアレイシステムにおいて、

前記各上位ファイバインタフェースに対応してそれぞれ設置されたキャッシュメモリ同士間でデータと管理情報を授受するための通信バスを有することを特徴とするディスクアレイシステム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明はファイバインタフェースを使用するコンピュータシステムに接続される磁気ディスク装置、磁気テープ装置、半導体記憶装置または光ディスク装置などの記憶媒体を制御する記憶システムに関する。

【０００２】

【従来の技術】ディスクアレイシステムは、ＲＡＩＤ（Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙｓｏｆ　Ｉｎｅｘｐｅｎｓｉｖｅ　Ｄｉｓｋｓ）とも呼ばれ、複数のディスク装置を列状（アレイ状）に配置した構成をとり、ホストからのリード要求（データの読み出し要求）およびライト要求（データの書き込み要求）をディスクの並列動作によって高速に処理するとともに、データに冗長データを付加することによって信頼性を向上させた記憶装置である。ディスクアレイシステムは、冗長データの種類と構成により５つのレベルに分類されている（論文：「Ａ　Ｃａｓｅ　ｆｏｒ　Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙｓ　ｏｆ　Ｉｎｅｘｐｅｎｓｉｖｅ　Ｄｉｓｋｓ　（ＲＡＩＤ）」，Ｄａｖｉｄ　Ａ．Ｐａｔｔｅｒｓｏｎ，Ｇａｒｔｈ　Ｇｉｂｓｏｎ，ａｎｄ　Ｒａｎｄｙ　Ｈ．Ｋａｔｚ，Ｃｏｍｐｕｔｅｒ　Ｓｃｉｅｎｃｅ　Ｄｉｖｉｓｉｏｎ　Ｄｅｐａｒｔｍｅｎｔ　ｏｆ　Ｅｌｅｃｔｒｉｃａｌ　Ｅｎｇｉｎｅｅｒｉｎｇ　ａｎｄ　Ｃｏｍｐｕｔｅｒ　Ｓｃｉｅｎｃｅｓ，Ｕｎｉｖｅｒｓｉｔｙ　ｏｆ　Ｃａｌｉｆｏｒｎｉａ　Ｂｅｒｋｅｌｅｙ）。

【０００３】ホストからのライト要求時には、ディスクアレイ内部では、ホストから受け取ったデータを特殊なアドレス変換によって分割し、一時的にキャッシュに格

納後、順次、ディスク装置へと並列に書込んでいく方式が一般的である。逆に、ホストからのリード要求時には、分割したデータをディスク装置から並列に読み出してキャッシュへと格納し、さらにキャッシュから読み出した分割データを元のデータとして束ね、ホストへと転送する。

【０００４】図２はＲＡＩＤ技術による従来ディスクアレイ構成図である。ディスクアレイシステムは、ホストインタフェース２００に接続するディスクアレイのファイバまたはＳＣＳＩの上位インタフェース２０２と、インタフェース２０２と内部バス２０３で互いに接続されるＭＰＵユニット２０４及びキャッシュコントローラ２０５と、キャッシュコントローラ２０５に接続するＨＤＤインタフェース２０７と、ＨＤＤ群２０９と、から構成される。

【０００５】図２では、上位インタフェースを１個として書いているが、複数であってもよい。この場合に、図２の構成において、共有するキャッシュメモリに複数の上位インタフェースが接続される構成となっており、上位インタフェースの転送能力を活かしきるため、上位インタフェースとキャッシュメモリ間のデータバスの転送能力が上位インタフェース転送能力以上であることが前提となる。ここで言う転送能力とは、単位時間あたりに送ることのできるデータ量のことであり、例えば、１秒間に１ギガバイトのデータを転送できる能力を１ギガバイト毎秒（ＧＢ／ｓ）と表すこととする。

【０００６】

【発明が解決しようとする課題】図２に示す従来アーキテクチャでは、ホストからのリード要求において、ディスクからキャッシュメモリにデータを送り、その後、キャッシュメモリからホストにデータを送るという流れである。ホストインタフェースの転送能力を１とした場合、ディスクインタフェースの転送能力は１、キャッシュメモリの転送能力は２となる。キャッシュメモリは、データの入力と出力があるため、ホストインタフェース、および、ディスクインタフェースの２倍の転送能力となる。ホストインタフェースの転送能力を生かしきるためには、キャッシュメモリの転送能力はホストインタフェースの２倍が必要となる。同様に、上位ホストの個数がＮ倍となれば、上位ホストインタフェースの転送能力もＮ倍となり、キャッシュメモリの転送能力は２Ｎ倍となる。

【０００７】このことからも分かるように、従来ディスクアレイのアーキテクチャにおいて、上位側の転送能力以上に、内部バスの転送能力を上げると共に、キャッシュメモリの転送能力も内部バス転送能力の向上に伴って上げるのが必須である。

【０００８】本発明の目的は、前述した従来技術の課題を解決し、複数の上位ファイバインタフェースを同時に駆使し、一つの高速ホストファイバインタフェースとデータやり取りを行うことにより、ホストファイバインタフェースに比べてその転送速度の低い内部バスを有するディスクアレイでも、転送速度が高速であるホストファイバインタフェースへの対応を可能とすることにある。

【０００９】また、本発明の他の目的は、ディスクアレイ内部で複数上位ファイバインタフェースを有する場合、ホストに複数ファイバインタフェースをアクセスする時の管理、或はターゲットＩＤ（複数ファイバのそれぞれのポートに付したアドレス）の切替負荷を掛けずに（ターゲットＩＤを切り替えた後にファイバ接続するのではホストに負荷が掛かるので）、一つのファイバインタフェースをアクセスするようにする手段を提供することにある。

【００１０】

【課題を解決するための手段】前記課題を解決するために、本発明は主として次のような構成を採用する。

【００１１】ホストファイバインタフェースからのデータを授受する複数の上位ファイバインタフェースと、前記複数上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライブインタフェースと、を備えたディスクアレイシステムであって、前記複数の上位ファイバインタフェースの各上位ファイバインタフェースに対応してそれぞれキャッシュメモリを備えているディスクアレイシステム。

【００１２】また、ホストファイバインタフェースからのデータを授受する複数の上位ファイバインタフェースと、前記複数上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライブインタフェースと、を備えたディスクアレイシステムであって、前記ホストファイバインタフェースと前記上位ファイバインタフェースとの間にデータ配分制御部を設け、前記データ配分制御部は、ライト時には、前記ホストファイバインタフェースからのデータを複数に分割して前記複数の上位ファイバインタフェースに供給するとともに、前記複数上位ファイバインタフェースを介して前記分割されたデータを並列に複数キャッシュメモリに同時に転送する、また、リード時においては、複数キャッシュメモリからのデータを並列に複数の上位ファイバインタフェースを通してデータ配分制御部に転送して前記データ配分制御部で分割されたデータを元のデータとして束ね、ホストへリードデータを返すディスクアレイシステム。

【００１３】

【発明の実施の形態】本発明の実施形態に係るディスクアレイシステムについて、図面を用いて以下説明する。図１は本発明の実施形態に係るディスクアレイシステムであり、ディスクアレイシステムは、図１に示すように、ホストファイバインタフェース１００にファイバケ

ーブル１０１を介して接続されるデータ配分制御部１０３と、前記データ配分制御部１０３にファイバケーブル１０８－０，１を介して接続される上位ファイバインタフェース１０２－０と１０２－１と、上位ファイバインタフェース１０２－０，１に接続される一時的にデータを保存するキャッシュメモリの制御コントローラ１０４－０と１０４－１と、これらのキャッシュコントローラ１０４－０，１配下のキャッシュメモリ１０５－０，１と、HDDインタフェース１０６と、ディスク群HDD１０７と、から構成される。キャッシュコントローラ１０４－０と１０４－１間でキャッシュ管理情報とデータの交信を出来るように、高速なデータバス１１１により接続される。

【００１４】前述したディスクアレイの構成において、ホストは、転送速度がディスクアレイのそれよりも高速であるファイバインタフェースを介してディスクアレイに対してライトする時、ディスクアレイのデータ配分制御部１０３で一旦データを受け取り、ディスクアレイ内部の上位ファイバインタフェースの数に応じてデータを分割する。また、受け取ったホストコマンドから上位ファイバインタフェースへのライトコマンドを生成し、各上位ファイバインタフェース１０２－０，１を介してキャッシュメモリ１０５－０，１にライトする。

【００１５】キャッシュメモリ１０５で分割されたデータブロックにホストデータを復元するための関連性を持たせた後、冗長情報をつけ、空き時間でHDD群１０７にライトする。また、ホストからのリード命令に対しては、キャッシュメモリ上にデータが存在する場合はキャッシュメモリ１０５から、キャッシュメモリにデータが存在しない時は、HDD群１０７より、データ配分制御部１０３に要求データを送り、データ配分制御部１０３で連続したデータにし、ホストに返す（詳細は後述）。

【００１６】次に、図３に基づいてデータ配分制御部１０３の機能乃至作用を説明する。図３に示す構成は、図１のデータ配分制御部１０３の詳細構造である。ホストファイバインタフェース１００と上位ファイバインタフェース１０２間の転送スピード差を十分に吸収するため、各上位ファイバインタフェースにおいて、２面のバッファを設け、総計４面とする。各バッファの容量は等しい。ホストファイバドライブ制御部３０１に４面のデータ転送バッファ３０４－０，１，２，３が、３０１のホストファイバと同速度を有する内部バス３０３、３０５－０，１，２，３とバッファバススイッチ３０２を介して接続され、また、前記の４面データバッファが上位ファイバドライブ制御部３０７と同等な転送速度を有する内部データバス３０８、３０９、３１１－０，１，２，３とバススイッチ３０６を介して上位のファイバドライブ制御部３０７－０，１に接続される。

【００１７】バッファスイッチ３０２によりホストファイバドライブ制御部３０１が接続したいバッファへの切り替えを行い、３０１が使用していないほかの３面バッファを上位ファイバドライブ制御部３０７－０，１に使わせることを可能とする。また、バススイッチ３０６の切り替えにより、４面バッファの内の２面が同時にファイバドライブ制御部３０７－０，１との間でデータ転送を行うことができる。前記バススイッチの切り替え動作は後述の転送リストに基づいて行われる。バススイッチを導入することにより、バス転送能力は勿論、バッファとファイバドライブ制御部間のデータ転送速度が最大限に引き出せる。

【００１８】MPUユニット部３１０にホストコンピュータからのコマンドを記憶するエリアを設ける。アクセス先のアドレス及び転送データ長をホストから受け取ったコマンドより割り出し、転送データ長を上位ファイバインタフェースの数で割った値が各上位ファイバドライブ制御部３０７－０，１が転送するデータの長さとなり、更にこの転送データ長を用いてデータ配分制御部内の一つバッファの大きさで割ったものを各上位ファイバドライブ制御部３０７－０，１が起動するバッファの回数とする。例えば、上位ファイバインタフェースの数はN、バッファのサイズはMとして、ホストからのリードまたはライトデータの長さはLとする場合は、各上位ファイバインタフェースによる転送データ長はL／Nとなり、各上位ファイバインタフェースが起動するバッファの回数は（L／N）／Mとなる。前記転送回数を用いてホストファイバドライブ制御部３０１、上位ファイバドライブ制御部３０７に転送リストを生成する。

【００１９】図８は上位ファイバドライブ制御部３０７における転送リストの例であり、上位ファイバドライブ３０７の制御部はこの転送リストに従ってバッファからキャッシュへデータ転送を行う。図８の（１）はファイバドライブ制御部３０７－０における制御部転送リストであり、図８の（２）はファイバドライブ制御部３０７－１における制御部転送リストである。

【００２０】また、２面のキャッシュメモリの管理テーブルのコピーをMPUユニット３１０に設け、キャッシュメモリにある管理テーブルが更新された場合は、その更新が上位ファイバインタフェースを通じてMPUユニット３１０にある管理テーブルに反映される。ホストからのライトコマンドに対し、バッファの容量ではなく、キャッシュ管理テーブルを検索し、キャッシュメモリの空き容量を見せることにより、ホストからの大容量なデータ転送要求を高速に応じることが可能となる。

【００２１】内部バッファの管理を単純化するために、バッファに連続したアドレスを持たせる。ホストファイバドライブ制御部３０１と上位ファイバドライブ３０７の間でバッファの使用衝突を回避するため、バッファ使用中で有るかどうかのフラグビットを設ける。バッファ使用中であれば、該当バッファのフラグビットを０（使用不可の印）にし、使用完了後（使用バッファ中のデー

タの出力後）、該当バッファのフラグビットを１（使用可の印）にし、ファイバドライブがバッファ使用可の状態とする。たとえば、ホストファイバドライブ３０１からバッファ１（３０４－１）にデータをライトする時はバッファ１のフラグが０となり、上位ファイバドライブ制御部３０７からはバッファ１へのアクセスができないが、バッファ２、３０４－２のフラグが１であれば、上位ファイバドライブからのリードまだはライトは可能である。

【００２２】ホストからのライト処理の流れを図６を用いて説明する。ここで、バッファ６０４－０，１，２，３の容量を３００ＫＢと仮定する。ホストからのライト要求は１．８ＭＢで、バッファ総容量より大きい場合の流れを示す。

【００２３】データ配分制御部にてホストからのライトコマンドを解読し、ライト先アドレスＬＢＡ（Ｌｏｇｉｃ Ｂｌｏｃｋ Ａｄｄｒｅｓｓ）と転送データ長ＬｅｎｇｔｈをＭＰＵユニット部６０９のメモリに一時保存する。同時にキャッシュの容量を確認して空き容量が要求データ長より大きい場合に、ホストに全数転送の応答を返し、一回で１．８ＭＢのデータを受け取る。キャッシュ容量が少なく、１．８ＭＢデータを一回で受け取れない場合はホストにキャッシュ空き容量分データ転送の要求を返す。また、キャッシュに空きができた時点で、残りのデータを受信する。図示する例では、典型的なキャッシュ空き容量が十分あるケースのみ紹介する。

【００２４】ホストからの１．８ＭＢのデータをホストファイバドライブ制御部６０１よりブッファ０に３００ＫＢ、バッファ１に３００ＫＢ、バッファ２に３００ＫＢ、バッファ３に３００ＫＢ、バッファ０に３００ＫＢ、バッファ１に３００ＫＢを格納していく。バッファ切替のタイムチャットを図４に示す。

【００２５】ホストファイバドライブ制御部６０１がバッファ０（６０４－０）へのデータ格納が完了したら、前記バッファ０のフラグを１にし、バッファ１（６０４－１）へのデータ格納を始める。それと同時に上位ファイバドライブ０（６０８－０）にホストからのライトアドレスＬＢＡと上位ファイバドライブ６０８による転送データ長Ｌｅｎｇｔｈ／２情報を含むライトコマンドを発行し、上位ファイバドライブ０（６０８－０）よりバッファ０（６０４－０）にあるデータをキャッシュ０（６１２－０）へ転送し始め（フラグを０にして）、転送完了後、バッファ０（３０４－０）のフラグビットを１に戻し、他のファイバドライブ制御部が使用できる状態にする。

【００２６】ホストファイバドライブ制御部６０１がバッファ１（６０４－１）へのデータ格納終了後、バッファ２（６０４－２）へのデータ格納を続けるが、それと同時に、前記と同様に、上位ファイバドライブ１（６０８－１）に６０８－０に発行したライトコマンドを発行し、ファイバドライブ１（６０８－１）よりバッファ１（６０４－１）のデータをキャッシュ１へ転送を行う。ホストファイバドライブ制御部６０１がバッファ２（６０４－２）へのデータ格納終了後、バッファ３（３０４－３）へのデータ格納を続け、それと同時に、上位ファイバドライブ０（６０８－０）がバッファ２（６０４－２）のデータを引き続きキャッシュ０へ転送する。キャッシュ０へ転送終了の後、バッファ２（６０４－２）のフラグビットを１にセットする。並行してホストファイバドライブ制御部より、バッファ０（６０４－０）へホストデータを入れ始める。ホストファイバドライブ制御部はバッファ３（６０４－３）へのデータ格納が終了後は、バッファ３のデータがファイバドライブ１（６０８－１）よりキャッシュ１へ転送される。

【００２７】ホストファイバドライブ制御部６０１が最後の３００ＫＢをバッファ１に格納すると同時に上位ファイバドライブ０（６０８－０）がバッファ０（６０４－０）のデータをキャッシュへ転送する。ホストファイバドライブがバッファ１へのデータ格納は完了後は、上位ファイバドライブ１（６０８－１）より、バッファ１のデータをキャッシュ１へ格納するという繰り返しになる。

【００２８】上記のように複数面バッファを使いまわすことにより、ホストからのデータ転送を中断せずに、ホストデータを９００ＫＢの二つのブロックに分割してキャッシュに格納することが可能となった。キャッシュがライト終了後に各上位インタフェースにライト完了の報告を上げ、データ配分制御部で各上位インタフェースからライト終了の信号が来たらホストにライト完了を報告する。

【００２９】上記方法の実現は、上位ファイバドライブ６０８のトータルな転送能力がホストファイバドライブ６０１の転送能力より大きいことにより保証される。ホストとデータ配分制御部間（図５の（１））、データ配分制御部と上位ファイバインタフェース間（図５の（２））のデータ転送タイムチャートを図５に示す。図５の（１）に示す上下平行線の左側にはホストファイバインタフェースにおける送信又は受信内容が経時的に上から順に記載されていて、その右側にはデータ配分制御部における受信又は送信内容が同様に記載されている。図５の（２）にはデータ配分制御部と上位ファイバインタフェース部との送信又は受信内容が図５の（１）と同様に経時的に記載されている。

【００３０】図６において上位ファイバドライブ０’制御部（６１０－０）配下のキャッシュ０（６１２－０）に分割されたデータブロック０、９００ＫＢが格納され、ファイバドライブ１’制御部（６１０－１）配下のキャッシュメモリ１（６１２－１）にはブロック１、９００ＫＢが格納される。ここで、図６に示す上位ファイバドライブ０’制御部６１０－０及び上位ファイバドラ

イブ１’制御部６１０－１は、図１に示す上位ファイバインタフェース１０２－０及び上位ファイバインタフェース１０２－１に対応するものである。

【００３１】キャッシュの管理を簡単化するため、２面のキャッシュをアドレス連続な一面のキャッシュメモリとして管理することにより、現在のキャッシュ制御アルゴリズムはそのまま流用可能である。キャッシュメモリ０，１にそれぞれキャッシュ管理テーブルＳＧＣＢ０，１（Ｓｅｇｍｅｎｔ　Ｃｏｎｔｒｏｌ　Ｂｌｏｃｋ）を設け、キャッシュ０のテーブルをマスタとし、かつ、キャッシュ０にキャッシュ１の管理テーブルのコピーを持たせる。キャッシュ上での管理単位はセグメントであり、ブロック０，１に２分割された１．８ＭＢのホストデータはさらに、セグメント単位でキャッシュ上で分割され、ＳＧＣＢ０，１より管理される。セグメントはキャッシュ上にある配置は連続でなくてもよい。ＳＧＣＢにはセグメントのアドレスを持っているほか、キャッシュにステージしたデータのＬＢＡ（Ｌｏｇｉｃ　Ｂｌｏｃｋ　Ａｄｄｒｅｓｓ）などの情報を有する。

【００３２】キャッシュにデータが書かれた後はＳＧＣＢ１の変化分をキャッシュ間のバスを介してキャッシュ０にあるＳＧＣＢ１のコピーに反映される。マスタＳＧＣＢ０が同一ＬＢＡを有するＳＧＣＢ１に属するセグメントを、ホストライトデータの内の一部とみなし、ＳＧＣＢは図７のように再構築される。よって、ホストデータが統合して管理される。図７に示す符号において、ＳＥＧＬのＳＥＧはセグメント（Ｓｅｇｍｅｎｎｔ）の意味であり、ＳＥＧＬのＬは左（Ｌｅｆｔ）の意味である。キャッシュが２面あって位置的に左側のキャッシュ０にあるＳＥＧがＳＥＧＬとし、右にあるキャッシュ１のＳＥＧがＳＥＧＲである。また、ＳＥＧＬ又はＳＥＧＲに続くＡＤＲはアドレスの意味である。

【００３３】図７に示すように、ホストからのデータが左又は右キャッシュに分割されるが、分割されたデータが各キャッシュでＳＥＧ単位で管理される。２面キャッシュのＳＥＧに図７に示す関係を持たせると、分散したデータが分散しているキャッシュで１つのデータとして管理、使用することが可能となる。

【００３４】キャッシュコントローラ０，１（６１２－０，１）にはデータ冗長性を付けする機能を持たせ、規格のＲＡＩＤ３又はＲＡＩＤ５の時はＨＤＤ群に保存のための冗長性コードをキャッシュ内で生成するときには、キャッシュ０，１（６１２－０，１）が同期を取りながら、冗長コードを生成する。キャッシュ０（６１２－０）でブロック０のパリティを算出後、そのパリティをと必要なデータをキャッシュ１（６１２－１）に渡して、前記キャッシュ１でブロック１のデータパリティを合成する。ＨＤＤインタフェース（図１の１０６）が冗長性のあるデータブロック０，１をＨＤＤ群（図１の１０７）への書き込みを担当する。

【００３５】データブロック０をＨＤＤ群（１０７）へ書きこみ先はＳＧＣＢ０が持っているＬＢＡとデータ転送長情報より取り決めれる。ブロック１のＨＤＤ群への書きこみ先はＳＧＣＢ１が持っているＬＢＡ＋冗長性のあるブロック０の長さとしたＬＢＡより取り決められる。ＨＤＤ群へデータをライトする際は、先にブロック０、次にブロック１との順で行われる。ＲＡＩＤ１の時は、キャッシュ間でデータコピーを行い、両キャッシュにブロック０、１のある状態に達した後、ＨＤＤインタフェースより各配下のＨＤＤへの書きこみを行う。

【００３６】次に、ホストからの書きこみ要求が２００ＫＢであって、バッファ総容量より小さい場合の流れについて述べる。ホストからの２００ＫＢのデータをホストファイバドライブ制御部６０１よりバッファ０（６０４－０）に１００ＫＢ、バッファ１（６０４－１）に１００ＫＢずつ格納し、バッファ２、３を使用せず。ホストファイバドライブ制御部６０１がバッファ０（６０４－０）へのデータ格納を完了後、バッファ０のフラグを１にし、バッファ０（６０４－０）を他ファイバドライブ制御部が使用可の状態に設置する。

【００３７】その後バッファ１（６０４－１）へのデータ格納の開始と同時に上位ファイバドライブ０（６０８－０）が前記バッファ０のデータを指定されたキャッシュ０（６１２－０）に転送し始め、転送完了後、バッファ０のフラグビットを１に戻し、他ファイバ制御部使用できる状態にする。ホストファイバドライブ制御部６０１がバッファ１（６０４－１）へのデータ格納が終了後、バッファ１を使用可の状態した後、同様に、上位ファイバドライブ１が前記バッファ１のデータを指定されたキャッシュ１（６１２－１）にデータ転送を行う。キャッシュ０，１（６１２－０，１）の配下にそれぞれ１００ＫＢのホストデータが格納される。その後の処理は前記１．８ＭＢの転送例と同様である。

【００３８】次に、リードヒット／ミス判定はＳＧＣＢ０のＬＢＡとセグメント情報より行われる。ホストからのリードコマンドがデータ配分制御部から、上位ファイバインタフェース０に送られる。要求ＬＢＡと要求データがキャッシュ上に存在する場合は、ヒットと判定し即座にデータ転送を始めることが可能である。要求ＬＢＡが存在するが、要求データの一部が既にＨＤＤに書きこまれて、そのデータが使用していたセグメントを他のデータに割り当てて、キャッシュ上にデータの一部を存在しない場合は、キャッシュヒットと判定するが、キャッシュ上に欠ける分のデータをＨＤＤよりリードしてからデータ転送を始める。ホストコマンドが要求するＬＢＡとＳＧＣＢ０が持っているＬＢＡ情報と一致しない場合はミスと判定する。

【００３９】ホストの要求により、前記に保存した１．８ＭＢデータをディスクアレイより読み出す時の流れを述べる。

【００４０】データ配分制御部でホストからのリードコマンドを解釈し、読み出し先、要求データ長を割り出して一時保存する。要求読み出しデータ長を上位ファイバチャンネルの数で割ったものを各上位ファイバドライブ制御部が転送するデータ長とする。そして、上位ファイバインタフェース０に読出先のアドレスＬＢＡ、転送データ長情報を含むリードコマンドを発行する。キャッシュ０で前記規則でヒット／ミスの判定をし、ヒットの場合はキャッシュ０，１からデータを返す。

【００４１】ミスの場合はリードアドレスＬＢＡ、データ転送長情報をＨＤＤインタフェース１０６に送り、１０６より、ホストデータの半分、ブロック０をキャッシュ０に送る。その後、キャッシュ０からキャッシュ１経由でＨＤＤインタフェース１０６にホストデータの半分、ブロック１をリードするアドレスＬＢＡ、伝送データ長情報を含むリードコマンドを発行する。ブロック０の転送完了後、１０６よりブロック１をキャッシュ１へ転送する。ブロック１のリードＬＢＡはブロック０のＬＢＡ＋転送データ長とする。

【００４２】キャッシュにリードデータを入れた後、データ配分制御部にデータを転送するために、前記ファイバドライブ制御部０，１に、キャッシュからのデータをバッファに入れる順番に関する図９の転送リストを生成する。データ配分制御部バッファに入れたキャッシュからのデータをホストに返すためにホストファイバドライブ制御部３０１にも図１０に示す形の転送リストを生成する。

【００４３】キャッシュ側のファイバドライブ制御部０’，１’（６１０－０，１）がデータ配分制御部からのリードコマンドを受けとり、ＨＤＤインタフェース０，１にリード先アドレス、リードデータ長を指示することにより、前記ＨＤＤインタフェース０，１はキャッシュ０にブロック０を、キャッシュ１にブロック１を転送する。その後、ファイバドライブ０’，１’より、データブロック０，１をデータ配分制御部へ送信する。上位ファイバドライブ０，１制御部（３０７－０，１）が受け取ったデータを図９の転送リストにしたがって３回づつデータ配分制御のバッファへデータ転送を行う。

【００４４】上位ファイバドライブ制御部０、１による一回目の転送では、バッファ０，１（３０４－０，１）が満杯になった後、上位ファイバドライブ制御部０，１が前記バッファ０，１を放し、バッファ使用状態のフラグが１にしてから、ホストファイバドライブ３０１が起動され、バッファ０，１の順からホストにデータを返す。その同時、上位ファイバドライブ０，１よりバッファ２，３へ２回目のデータ転送を行い、完了後、該当バッファのフラグビットを１にしてホストファイバドライブ制御部が使用可能の状態にしておく。バッファ０，１からホストへのデータ転送が終わり、使用可能状態となると、再びバッファ０，１に、上位ファイバドライブ制御部０，１より最後の６００ＭＢを格納される。それと同時に、ホストファイバドライブ制御部３０１がバッファ２，３よりホストへデータを返し、最後の６００ＭＢをバッファ０，１よりホストへ転送する。

【００４５】続いて、前記例で保存した２００ＭＢデータをＨＤＤ群より読み出し時の流れを述べる。データ配分制御部２０３でホストからのリードコマンドを解釈し、読み出し先、要求データ長を割り出して一時保存する。要求読み出しデータ長を上位ファイバインタフェース（６０８－０，１）の数で割ったもの、１００ＭＢを上位ファイバドライブが転送するデータ長とし、上位ファイバドライブ制御部０，１の制御部に、バッファ０，１のみ一回ずつ使用する。データ配分制御部は上位ファイバインタフェース０に読出先のアドレスＬＢＡ、転送データ長情報を含むリードコマンドを発行する。

【００４６】キャッシュ０で前記規則でヒット／ミスの判定をし、ヒットの場合はキャッシュからデータを返す。ミスの場合はリードアドレスＬＢＡ、データ転送長情報をＨＤＤインタフェース１０６に送り、１０６より、ＨＤＤから１００ＭＢのデータブロック０をキャッシュ０、ブロック１をキャッシュ１に格納する。キャッシュ０，１から各バッファにおいて１００ＭＢのデータを送信のための転送リストが生成される。データ配分制御部バッファにキャッシュからのデータをホストに返すためにホストファイバ制御部と上位ファイバインタフェース制御部に、図９、図１０と同様な転送リストを生成する。

【００４７】前記転送リストにしたがって、上位ファイバドライブ制御部０，１はデータ配分制御のバッファ０，１へ同時に１００ＭＢのデータを転送する。バッファ０，１がホストファイバドライブ制御部が使用できる状態になった後はバッファ０，１から１００ＭＢずつのデータをホストに返す。

【００４８】本実施形態にあるＨＤＤ群はファイバインタフェースを有するＨＤＤ採用しており、ＨＤＤインタフェースにＨＤＤ群を制御するためのファイバドライブＦＰＣが使用される。パリティ生成機能をキャッシュコントローラに持たした形となっているが、キャッシュコントローラではなく、ＨＤＤインタフェースにパリティ生成機能を持たすことも可能である。

【００４９】また、本実施形態において、ディスクアレイ内部に二つの上位ファイバーインタフェースを使用する例を挙げているが、上位ファイバインタフェースの数は二つに限らず、複数であればよい。また、本実施形態において、データ配分制御部とキャッシュ間はファイバチャンネルを介して接続されるが、直接内部バスで接続される場合も、同様な効果は得られる。

【００５０】以上説明したように、本発明の実施形態は、次のような構成、機能乃至作用を奏するものを含むものである。即ち、ホストファイバインタフェースから

のデータを授受する複数の上位ファイバインタフェースと、前記複数上位ファイバインタフェースに接続されたキャッシュメモリと、データを書き込み又は読み出す複数のディスク記憶手段と、前記記憶手段を制御するディスクドライブインタフェースと、を備えたディスクアレイシステムにおいて、前記ホストファイバインタフェースと前記上位ファイバインタフェースとの間に設けられたデータ配分制御部は、ホストファイバインタフェースからのデータを上位インタフェースの数に応じて複数ブロックに分割し、データ配分制御部に接続される複数の上位ファイバインタフェースを介して前記分割されたデータを並列に複数からなるキャッシュメモリに同時に転送を行うことと、前記ディスクアレイ内部が複数ファイバインタフェースがそれぞれのキャッシュメモリを持つことにより、上位ファイバインタフェースに繋ぐディスクアレイ内部バスの転送速度（即ち、図１に示す各内部バス１０９－０又は１０９－１の転送速度）がホストファイバインタフェースに比べてその転送速度が低くても転送速度が高速であるホストファイバインタフェースへの対応を可能とすることである。

【００５１】また、本実施形態は、データ配分制御部は自身が持つファイバインタフェースのＩＤをホストに見せ、ホストからのアクセス要求を一旦、バッファで受け取り分割し、分割したアクセス要求を同時に上位インタフェースに発行して実行させることにより、ホストがディスクアレイ中の複数ファイバインタフェースの存在を意識させないこととなる。換言すると、ファイバループ又はファイバスイッチによる接続と異なり、ホストファイバインタフェースがバッファを噛まして複数の上位ファイバチャンネルと接続し、ホストファイバインタフェースは上位ファイバインタフェースと直接に接続されないこととなる。

【００５２】

【発明の効果】本願において開示される発明の内、体表的なものによって得られる効果を簡単に説明すれば、下記の通りである。

【００５３】ディスクアレイ内部の複数の上位ファイバインタフェースを同時に動作させることより、ホストファイバインタフェースに比べてその転送速度の遅い内部バスを有するディスクアレイシステムでも、転送速度が高速であるホストファイバインタフェースを有するホストからの高速アクセスに速度を損うことなく、実行することが可能となった。

【００５４】また、データ配分制御部のバッファにおいて、ホストライト要求にキャッシュの空き容量で答えることと、複数面の小容量なバッファを使いまわしすることにより、小容量なバッファでホストとディスクアレイ間との大容量なデータ転送を可能にした。

【００５５】また、多面キャッシュを一元管理することで、現用キャッシュ制御マイクロに大きな変更を加えずに流用することが可能である。

【図面の簡単な説明】

【図１】本発明の実施形態に係るディスクアレイシステムの構成の概要を示す概念図である。

【図２】ＲＡＩＤ技術における従来のディスクアレイシステムの概略図である。

【図３】本実施形態のデータ配分制御部の構成を示す図である。

【図４】本実施形態におけるライト時のデータ配分制御部でのバッファ動作のタイミングを示す図である。

【図５】本実施形態におけるライト時のデータ配分制御部動作と、ホスト又は内部上位ファイバインタフェース間動作とのタイムチャート図である。

【図６】本実施形態におけるデータ配分制御部とキャッシュ間の構成を示す図である。

【図７】本実施形態におけるキャッシュ管理テーブルＳＧＣＢ０，１及びセグメントのキャッシュへの格納イメージ図である。

【図８】本実施形態におけるキャッシュライト時の上位ファイバドライブ転送用転送リストを示す図である。

【図９】本実施形態におけるキャッシュリード時の上位ファイバドライブ転送用転送リストを示す図である。

【図１０】本実施形態におけるリード時のホストファイバドライブ転送用転送リストを示す図である。

【符号の説明】

１００　ホストファイバインタフェース
１０１　ファイバケーブル
１０２－０～１　上位ファイバインタフェース
１０３　データ配分制御部
１０４－０～１　キャッシュコントローラ
１０５－０～１　キャッシュメモリ
１０６　ＨＤＤインタフェース
１０７－０～３　ＨＤＤドライブ
１０８－０～１　ファイバケーブル
１０９－０～１　内部バス
１１０　ファイバケーブル
１１１　内部キャッシュ間バス
２００　ホストインタフェース
２０１　Ｆｉｂｒｅ／ＳＣＳＩケーブル
２０２　上位インタフェース
２０３　内部バス
２０４　ＭＰＵユニット
２０５　キャッシュコントローラ
２０６　キャッシュメモリ
２０７　ＨＤＤインタフェース
２０８　Ｆｉｂｒｅ／ＳＣＳＩケーブル
２０９－０～３　ＨＤＤドライブ
３０１　ホストファイバドライブ制御部
３０２　バッファバススイッチ
３０３　内部バス

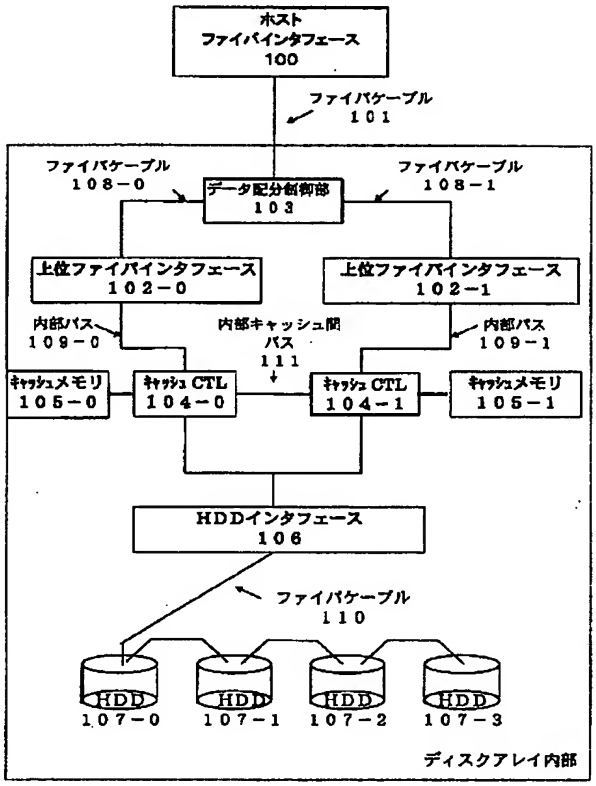
15

３０４－０～３　データバッファ
３０５－０～３　内部バス
３０６　バッファバススイッチ
３０７－０～１　上位ファイバドライブ制御部
６０１　ホストファイバドライブ制御部
６０２　バッファバススイッチ
６０３　内部バス
６０４－０～３　データバッファ
６０５　バッファバススイッチ

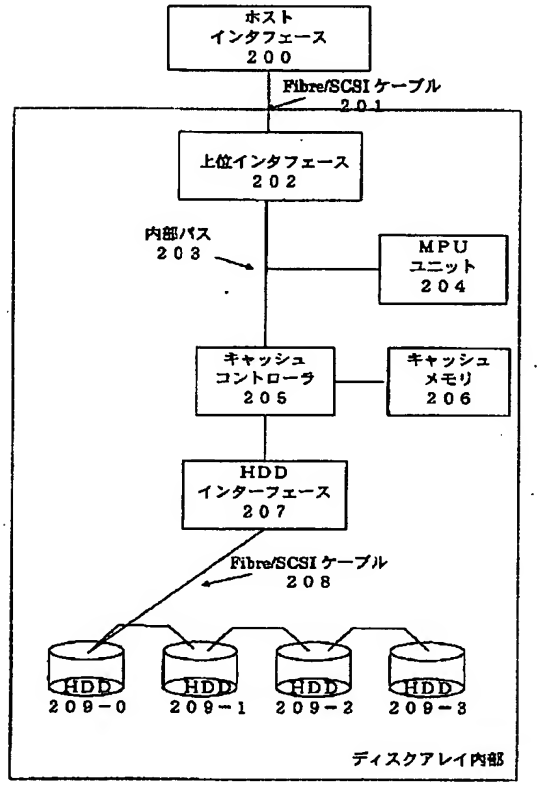
16

６０６～７　内部バス
６０８－０～１　上位ファイバドライブ０，１制御部
６０９　ＭＰＵユニット部
６１０－０～１　上位ファイバドライブ０’，１’制御部
６１１－０～１　ファイバケーブル
６１２－０～１　キャッシュコントローラ及びキャッシュメモリ
６１３－０～１　内部バス

【図１】



【図２】



【図９】

上位ファイバドライブ制御部転送リスト

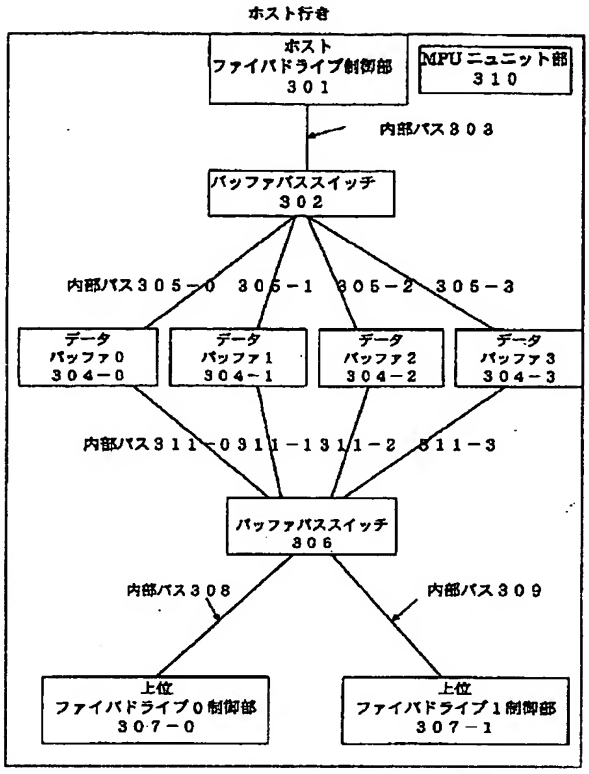
転送リスト	送信ファイバドライブ	送信先	読み出し先	転送データ長
1	0	バッファ０	ブロック０：SEGRxx	計３００MB
	1	バッファ１	ブロック１：SEGLxx.	計３００MB
2	0	バッファ２	ブロック０SEGRxx	計３００MB
	1	バッファ３	ブロック１SEGLxx	計３００MB
3	0	バッファ０	ブロック０SEGRxx	計３００MB
	1	バッファ１	ブロック１SEGLxx	計３００MB

【図１０】

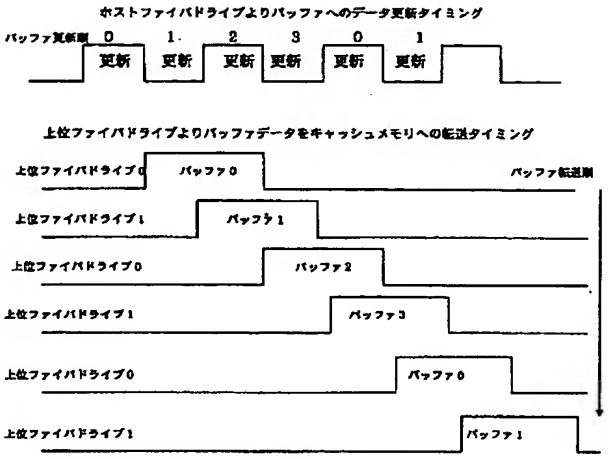
高速ホストファイバドライブ制御部転送リスト

ホストファイバドラバ転送リスト	読み出し先	読み出しデータ長
0	バッファ０	全領域
1	バッファ１	全領域
2	バッファ２	全領域
3	バッファ３	全領域
4	バッファ０	全領域
5	バッファ１	全領域

【図３】



【図４】



【図８】

（１）

ファイバドライブ３０７－０における制御部転送リスト

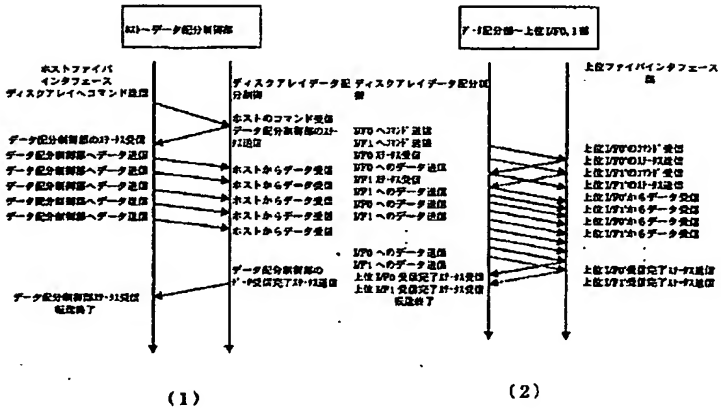
転送リスト	３０７－０	読み出し先	転送先
1	起動	バッファ０ ３０４－０	キャッシュ０
2	起動	バッファ２ ３０４－２	キャッシュ０

（２）

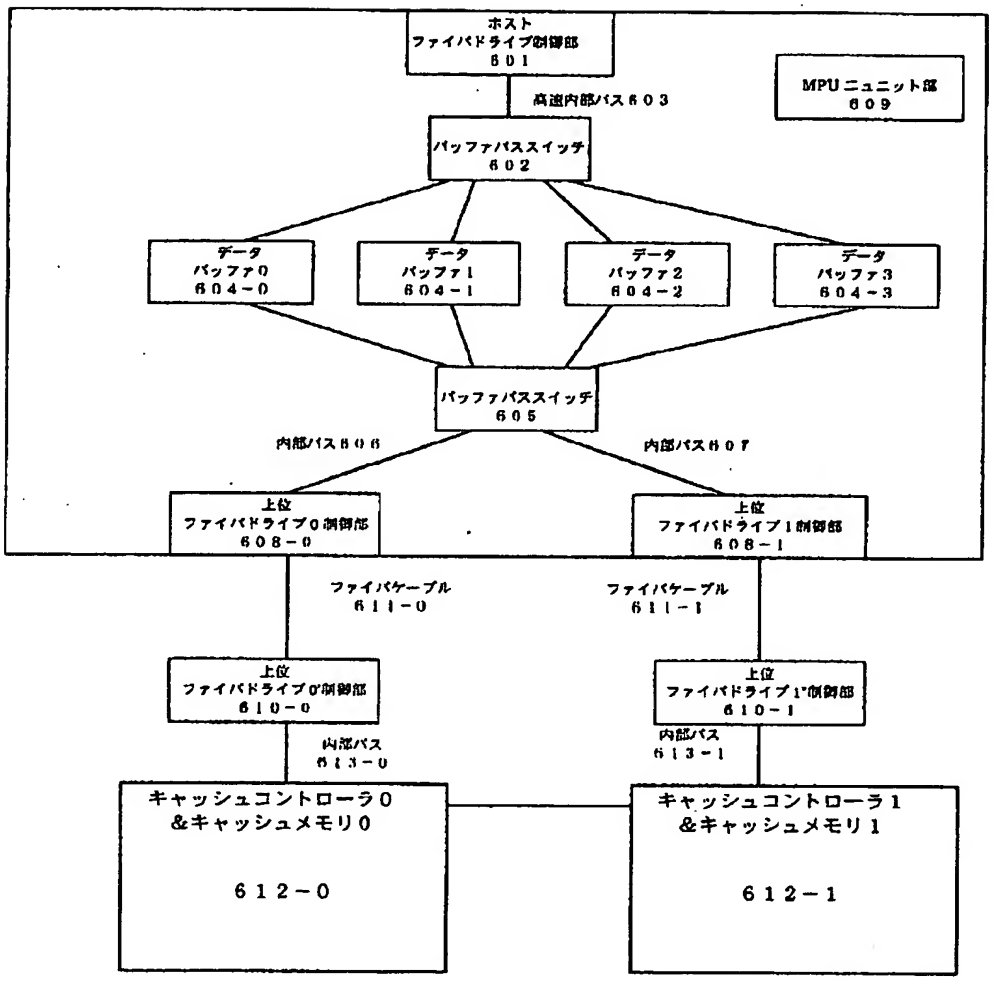
ファイバドライブ３０７－１における制御部転送リスト

転送リスト	３０７－１	読み出し先	転送先
1	起動	バッファ１ ３０４－１	キャッシュ１
2	起動	バッファ４ ３０４－３	キャッシュ１

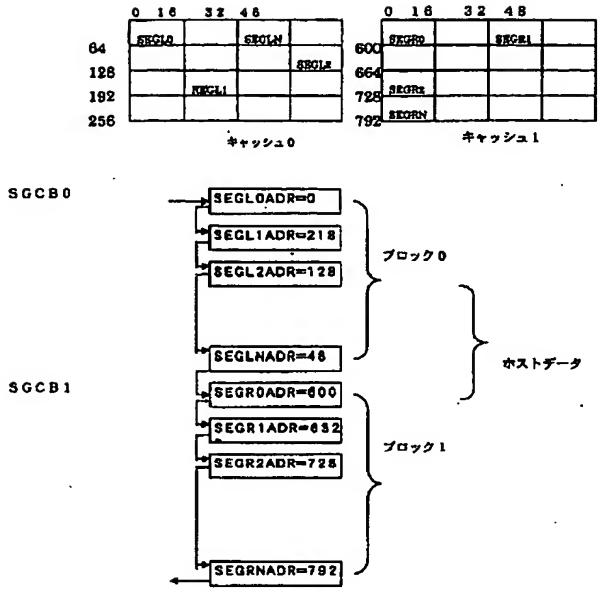
【図５】



【図６】


ホスト
ファイバドライブ制御部
６０１
MPUユニット部
６０９
高速内部バス６０３
バッファパススイッチ
６０２
データ
バッファ０
６０４－０
データ
バッファ１
６０４－１
データ
バッファ２
６０４－２
データ
バッファ３
６０４－３
バッファパススイッチ
６０５
内部バス６０６
内部バス６０７
上位
ファイバドライブ０制御部
６０８－０
上位
ファイバドライブ１制御部
６０８－１
ファイバケーブル
６１１－０
ファイバケーブル
６１１－１
上位
ファイバドライブ０'制御部
６１０－０
上位
ファイバドライブ１'制御部
６１０－１
内部バス
６１３－０
内部バス
６１３－１
キャッシュコントローラ０
＆キャッシュメモリ０
６１２－０
キャッシュコントローラ１
＆キャッシュメモリ１
６１２－１

【図７】



フロントページの続き

(72)発明者　八木沢　育哉
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

Ｆターム(参考)　5B005 JJ12 MM12 NN12
5B065 BA01 CA30 CC08 CE12 CE14 CE26